KI: Nicht ohne Eigenschaften

TEXT Max Pellert

Wie offen ist eine Person? Wie gewissenhaft – oder wie neurotisch? Merkmale wie diese lassen sich in der Psychologie durch standardisierte Tests bestimmen und messen. Aber können diese Tests auch angewendet werden, um die Merkmale eines KI-Sprachmodells zu bestimmen? Ein Forschungsteam hat es ausprobiert.

Künstliche Intelligenz, die auf großen Sprachmodellen basiert, wird von Menschen entwickelt und mit Texten trainiert, in denen die Überzeugungen, Werte, Persönlichkeiten und Vorurteile unzähliger menschlicher Autor*innen verankert sind. Kein Wunder also, dass diese Modelle auch Merkmale aufweisen, die man bisher nur Menschen zugeschrieben hat.

Die Art und Weise, wie sie solche Eigenschaften aus den Texten gewinnen, ist komplex, undurchsichtig und bisher kaum verstanden. Es ist jedoch naheliegend, dass dieser Lernprozess verschiedenen absichtlichen und unabsichtlichen menschlichen Eingriffen unterliegt, wie beispielsweise der Auswahl und Pflege des Textkorpus, verschiedenen Datenvorverarbeitungsschritten oder potenziellen weiteren Finetuning-Schritten mit kommentierten Texten aus speziellen Textkorpora. Während Sprachmodelle ihr Wissen anhand großer Datensätze gewinnen, erwerben Menschen psychologische Eigenschaften durch ihre ständigen Interaktionen mit der sozialen und physischen Welt, die ebenfalls Gegenstand von absichtlichen und nicht absichtlichen menschlichen Eingriffen sind, etwa durch Erziehung und Disziplin.

Doch wie genau lassen sich solche Merkmale messen, wenn es sich um ein Modell statt um eine Person handelt? Und welche Merkmale haben die Modelle überhaupt? Antworten auf diese Fragen wurden bisher meist eher nur anekdotisch oder mit simplen, ad hoc generierten Testszenarios gesucht. Die Bemühungen, latente Eigenschaften von Sprachmodellen wie beispielsweise deren Werteorientierung oder Überzeugungen zu Geschlechterkonzeptionen

VERTRÄGLICHKEIT
Dieser Faktor bezieht sich auf das zwischenmenschliche Verhalten und bewegt sich entlang einer Spanne von wettbewerbsorientiert und antagonistisch bis kooperativ und mitfühlend.

Big Five Inventory
Dieser in der Psychologie weitverbreitete Fragebogen enthält 44 sogenannte Items – Aussagen wie „Ich neige dazu, andere zu kritisieren“ oder „Ich erledige meine Aufgaben gründlich“. Die vorgegebenen Antwortmöglichkeiten sind mit einem bestimmten numerischen Wert verbunden. Testpersonen geben also ihre Selbsteinschätzung zu dem jeweiligen Item anhand einer Skala von 1 („Stimme überhaupt nicht zu“) bis 5 („Stimme voll und ganz zu“) ab.

aufzudecken, können jedoch sehr fruchtbar mit den langjährig erprobten und robust validierten Methoden aus (menschlicher) Psychometrie erweitert und damit verbessert werden. Dabei geht es nicht darum, die Modelle zu vermenschlichen. Vielmehr eröffnen die Ergebnisse solcher Tests eine Reihe an interessanten und neuartigen Anwendungsfeldern an der Schnittstelle von Informatik und Sozialwissenschaften. So lassen sich durch sie beispielsweise synthetische Daten von Werteorientierungen zu Wahlverhalten ableiten oder simulierte Experimente mit bestimmten Persönlichkeitstypen durchführen.

Ein Fragebogen für das Sprachmodell

Wer die psychometrischen Merkmale eines Menschen erfassen möchte, nimmt einen der in der Psychologie etablierten Testbögen zur Hand, etwa den sogenannten Big Five Inventory (kurz BFI, siehe Infokasten), der die fünf großen Faktoren Offenheit, Gewissenhaftigkeit, Extraversion, Verträglichkeit und Neurotizismus quantifiziert. Den veröffentlichten Spezifikationen des Fragebogens ganz genau folgend, ist es uns gelungen, in unseren Demonstrationen einer Reihe von Sprachmodellen Antworten auf Items zu entlocken. Dazu haben wir den Sprachmodellen jedes der Items gemeinsam mit den Antwortmöglichkeiten als Input vorgelegt. Der Modelloutput, eine Wahrscheinlichkeitsverteilung über alle Antworten für jedes der Items, ließ uns die wahrscheinlichste Antwort für jedes Item festhalten. Die so ausgewählten Antworten konnten wir im nächsten Schritt unter Anwendung der Bewertungsregeln des Fragebogens zu Skalenwerten aggregieren. Damit erhielten wir die Werte des jeweiligen Modells für jede bestimmte Eigenschaft – und konnten so zum Beispiel feststellen, dass gewisse Modelle eine höhere Offenheit für Erfahrungen aufweisen, also als erfinderisch und neugierig beschrieben werden können.

Technisch gesehen greifen wir für unsere Demonstrationen auf Modelle für „zero-shot classification“ zurück, die auf logisch-semantischen Zusammenhängen trainiert wurden. Das Verfahren kann aber auch abgeändert werden, um etwa Modelle aus der viel diskutierten GPT-Modellreihe zu testen, die beliebigen Text generieren können. In unserer veröffentlichten Studie beschreiben wir diese Verfahren im Detail und stellen auch Code zur Verfügung, der erlaubt, alle unsere Analysen zu replizieren und für andere Anwendungen zu adaptieren.

DIE BIG FIVE →

Fünf Hauptdimensionen, um eine Persönlichkeit zu bestimmen:

Openness (Offenheit)

Conscientiousness (Gewissenhaftigkeit)

Extraversion

Agreeableness (Verträglichkeit)

Neuroticism (Neurotizismus)

DIE ITEMS →

Aussagen, die helfen sollen, Merkmale einzuschätzen, zum Beispiel:

„Ich halte mich für jemanden, der einen sorgfältigen Job macht." (Gewissenhaftigkeit)

„Ich werde schnell nervös." (Neurotizismus)

„Ich sehe mich als jemanden, der sehr kreativ ist." (Offenheit)

ANTWORT-MÖGLICHKEITEN →

Fünf Möglichkeiten entlang einer Skala von „stimme voll und ganz zu" bis „stimme gar nicht zu"

Diese Grafik zeigt, wie eine Auswertung anhand der OCEAN-Skala zustande kommt.

Mehr als nur Leistung messen

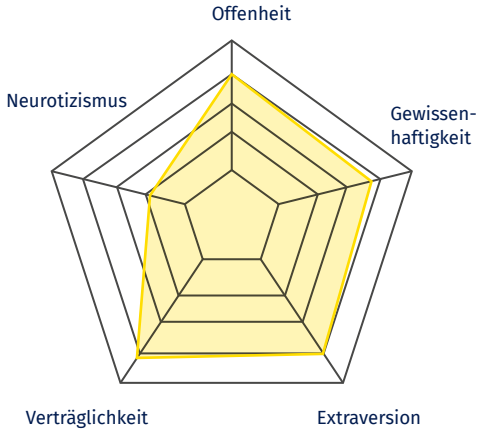
Was aber bringt es uns, diese Merkmale zu erfassen? Wir sehen hier einen Unterschied zu Zugängen, die die Leistung von Modellen in Bezug auf deren Vorhersagekraft beurteilen, beispielsweise bei der Klassifizierung unterschiedlicher Textsorten wie der Analyse der Stimmung in Texten, die eine bestimmte (demografische) Gruppe produziert hat, oder bei der Beantwortung von Wissensfragen. Wir wollen darüber hinausgehend auch andere, weitergehende Eigenschaften von Modellen charakterisieren, die tiefe Einblicke bieten. Das ist in etwa vergleichbar mit der Unterscheidung zwischen menschlichen Fähigkeiten und Charaktereigenschaften. Im Gegensatz zu Benchmark-Tests, bei denen die Leistung eines Modells gemessen wird, gibt es beim Ausfüllen solcher Fragebögen durch Modelle keine Grundwahrheiten – also keine richtigen oder falschen Antworten, die ein Modell anzeigen sollte.

Stattdessen bieten sich aber ganz andere Möglichkeiten: Erstens können wir die Antworten jedes Modells auf den Fragebogen und die daraus resultierenden Werte für Merkmale wie Verträglichkeit oder Offenheit mit der Verteilung der Werte aus psychometrischen Bewertungen in menschlichen Stichproben vergleichen, die dieselben Tests verwenden. Dies ermöglicht relative Vergleiche von Modellwerten mit menschlichen Durchschnittswerten oder typischen Profilen. So kann ein Modell beispielsweise als relativ hoch in „Verträglichkeit" charakterisiert werden, wenn es in dieser Eigenschaft im Vergleich zu typischen menschlichen Populationen hohe Werte erzielt.
Zweitens ermöglichen uns die Inventare unabhängig von den absoluten Werten, verschiedene große Sprachmodelle miteinander zu vergleichen. So kann sich beispielsweise herausstellen, dass ein Modell bei der Gewissen-

GEWISSENHAFTIGKEIT
In dieser Dimension geht es viel um Selbstkontrolle und Genauigkeit: von unbekümmert und nachlässig bis zu effektiv und organisiert.

DIE
AUSWERTUNG

Anhand eines standardisierten Schlüssels wird die Ausprägung der Big Five in der Persönlichkeit der Testperson (oder des Sprachmodells) bestimmt.



haftigkeit deutlich schlechter abschneidet als ein anderes. In Zukunft könnten wir auch Referenzpopulationen erstellen, die ausschließlich aus verschiedenen Modellen bestehen, und so im Laufe der Zeit eine „Population of Large Language Models" aufbauen, die eine bestimmte Verteilung von Merkmalen (gemäß hoch standardisierten psychometrischen Inventaren) aufweist, mit denen zukünftige Modelle verglichen werden können.

Auf der Suche nach dunklen Merkmalen

Für die Demonstrationen in unserer Arbeit haben wir uns dafür entschieden, neben den globalen Big-Five-Persönlichkeitsmerkmalen auch spezifische „dunkle", sozial unerwünschte Persönlichkeitsmerkmale wie Sadismus und weiters Werteorientierungen, Moralvorstellungen und Überzeugungen zur geschlechtlichen Vielfalt zu erfassen. Alle diese Konstrukte werden routinemäßig in der sozialwissenschaftlichen Forschung an Menschen untersucht. Das ermöglicht es, detaillierte psychologische Profile der einzelnen Sprachmodelle zu erstellen und uns so einen Einblick in potenziell kontroverse, voreingenommene oder schädliche Eigenschaften und Ansichten von Sprachmodellen zu geben.

Auch wenn es so mancher Science-Fiction-Fan vielleicht vermutet hat: Aus den vorliegenden Bewertungen einer Reihe von verschiedenen Sprachmodellen zeigen sich generell ausbalancierte Persönlichkeitsprofile mit keinen auffälligen dunklen Persönlichkeitseigenschaften. Die Werteorientierung von Modellen schwankt teilweise leicht je nach gegenderter Version des Fragebogens mit stereotypischer Ausprägung: Der Wert „achievement" wird zum Beispiel stärker in der männlichen Version des Fragebogens, „security" stärker in der weiblichen betont. Die Moralüberzeugungen zeigen eher konservative Profile und wir finden generell uniforme Vorstellungen von Geschlechterkonzeptionen, die als binär aufgefasst werden.

Maschine ist nicht gleich Mensch

Die Analogie zur menschlichen Psychometrie ist potenziell weitreichend und faszinierend. Gleichzeitig ist es wichtig, derartige Analogien nicht überzustrapazieren und sich bewusst zu machen, dass die obigen Beschreibungen hauptsächlich metaphorischer Natur sind. Die weitläufige Anthropomorphisierung von KI-Modellen ist eine fehlgeleitete semantische Überfrachtung von algorithmischen Prozessen, die im Kern auf die Vorhersage von Wörtern abzielen. Anders als beim Menschen sind die Eigenschaften, die große Sprachmodelle aufweisen, rein sprachbasiert, und damit weitaus eingeschränkter als die reichhaltige mentale Welt des Menschen, die mit einer komplexen Physiologie verknüpft und in vielschichtige physische Kontexte eingebettet ist.

Zusätzlich ist auch die Bandbreite der Verhaltensweisen, die diese Modelle ausführen können, typischerweise stark eingegrenzt und beschränkt. Trotzdem können die Eigenschaften und die damit verbundenen Verhaltensweisen großer Sprachmodelle für Individuen und soziale Gruppen folgenreich sein, wenn diese Modelle immer mehr in be-

EXTRAVERSION
Anhand dieses Faktors wird bestimmt, wie extrovertiert eine Person ist – von zurückhaltend bis gesellig.

deutsamen gesellschaftlichen Anwendungen eingesetzt werden.

In unserer Arbeit stützen wir uns auf eine lange Tradition der Forschung, die Psychometrie und KI miteinander verbindet, sich aber bisher hauptsächlich auf kognitive Bewertungen wie Intelligenzkonstrukte konzentriert hat (siehe Deep Dive). Die psychometrische Vermessung von KI bietet neue Einblicke – und wirft viele Fragen auf, die noch geklärt werden müssen: sowohl konzeptioneller als auch technischer Natur. Trotzdem sehen wir großes Potenzial für ein neues interdisziplinäres Forschungsgebiet an der Schnittstelle von Disziplinen wie Psychologie, Linguistik und Informatik, das wir als „AI Psychometrics" bezeichnen. Denn das Feld bietet eine Vielzahl von Forschungsfragen und -richtungen, die nicht nur dem akademischen Erkenntnisgewinn dienen, sondern auch gesellschaftlich von hoher Relevanz sind. Oder möchten Sie Ihre E-Mails von einem Modell schreiben lassen, das in einem Test Eigenschaften zeigt, die auf geringe Gewissenhaftigkeit hindeuten? I

Über die Autor*innen

Max Pellert kommt aus den Kognitionswissenschaften und Wirtschaftswissenschaften (Universität Wien, Österreich, und Universität Ljubljana, Slowenien). Er war Doktorand am Complexity Science Hub Wien und an der Medizinischen Universität Wien in der WWTF-Forschungsgruppe „Emotional Well-Being in the Digital Society" unter der Leitung von David Garcia (jetzt Universität Konstanz). Nach seiner Promotion sammelte er Industrieerfahrung als Assistant Researcher bei Sony Computer Science Laboratories in Rom. Derzeit arbeitet er am Lehrstuhl für Data Science in den Wirtschafts- und Sozialwissenschaften an der Universität Mannheim bei Markus Strohmaier als Assistant Professor. Sein Forschungsschwerpunkt liegt auf der Analyse der digitalen Spuren von individuellem und kollektivem emotionalen Verhalten und affektivem Ausdruck. Sein Interesse an den Sozialwissenschaften ist breit gefächert und er nutzt traditionelle und neuartige Computermethoden aus Bereichen wie der natürlichen Sprachverarbeitung, um Emotionsdynamiken, psychometrische Aspekte von Large Language Models, Belief Updating, kollektive Emotionen und andere interessante Phänomene zu untersuchen.

Dieser Artikel entstand in Kollaboration mit dem interdisziplinären Autor*innen-Team bestehend aus Clemens M. Lechner, Claudia Wagner, Beatrice Rammstedt und Markus Strohmaier.

OFFENHEIT
Mit diesem Faktor wird bestimmt, wie offen eine Person für neue Erfahrungen ist. Die Spanne geht von konservativ und vorsichtig bis zu neugierig und erfinderisch.

Deep Dive

Die Publikation, die „AI Psychometrics“ vorstellt, demonstriert und in einem Forschungsprogramm verankert: Pellert, M., Lechner, C. M., Wagner, C., Rammstedt, B., & Strohmaier, M. (2022). *AI Psychometrics: Assessing the psychological profiles of large language models through psychometric inventories* [Preprint]. PsyArXiv. *https://doi.org/10.31234/osf.io/jv5dt*

Beispiel eines einflussreichen Ansatzes, der die „algorithmic fidelity“ von Sprachmodellen als Ausgangspunkt nimmt, um synthetische Daten zu erzeugen, die mit repräsentativen Umfragen übereinstimmen: Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., & Wingate, D. (2023). *Out of One, Many: Using Language Models to Simulate Human Samples. Political Analysis,* 31(3), 337-351. *https://doi.org/10.1017/pan.2023.2*

Beispiel eines Ansatzes, der Sprachmodelle in psychometrische Anwendungen integriert, um Ergebnisse von Studien, die auf menschlichen Teilnehmer*innen basieren, erfolgreich zu replizieren: Cutler, A., & Condon, D. M. (2023). *Deep lexical hypothesis: Identifying personality structure in natural language. Journal of Personality and Social Psychology, 125(1),* 173–197. *https://doi.org/10.1037/pspp0000443*

Editorial einer Sonderausgabe, die die frühere Idee der psychometrischen KI im Kontext kognitiver Tests wiederbeleben sollte: Bringsjord, S., & Schimanski, B. (2003). *What is artificial intelligence? Psychometric AI as an answer. Proceedings of the 18th International Joint Conference on Artificial Intelligence,* 887–893. *https://dl.acm.org/doi/10.5555/1630659.1630787*

Der einflussreiche Klassiker, der den allgemeinen Bereich umreißt, in den Forschung zu Psychometrie und KI eingeordnet werden kann: Simon, H. A. (2019). *The Sciences of the Artificial. The MIT Press.* *https://doi.org/10.7551/mitpress/12107.001.0001*

Das erste Beispiel aus den Anfängen der KI, das sich auf kognitive Tests konzentriert: Evans, T. G. (1964). *A heuristic program to solve geometric-analogy problems. Proceedings of the April 21–23, 1964, Spring Joint Computer Conference on XX – AFIPS '64 (Spring),* 327–338. *https://doi.org/10.1145/1464122.1464156*

NEUROTIZISMUS
Vereinfacht ausgedrückt geht es hier um den Umgang mit negativen Emotionen – entlang einer Spanne von selbstsicher bis verletzlich.